Э. Б. Ершов

Выбор регрессии, максимизирующий несмещенную оценку коэффициента детерминации

Получена форма несмещенной оценки коэффициента детерминации для линейного уравнения регрессии, вычисляемая по выборочным данным из многомерного нормального распределения. Эту оценку предлагается применять как альтернативный критерий выбора факторов в регрессии.

1. Введение

Общеизвестен вариант исходных предположений метода наименьших квадратов (МНК), при котором используемые значения объясняемой переменной y и факторов $x_1,\dots,x_m$ в регрессии порождаются выборкой из многомерного невырожденного нормального распределения. Вместо его неизвестных параметров используются их оценки, вычисляемые по выборочным данным.

Нормально распределенная случайная величина y, получаемая при фиксированных значениях факторов $x_1,\dots,x_m$, представима в виде

$$y = a_0 + \sum_{j=1}^{m} x_j a_j + \varepsilon, \tag{1}$$

где коэффициенты $a_0,\dots,a_m$ — известные функции от параметров закона распределения случайной величины $(y, x_1,\dots,x_m)$ и ε — нормально распределенная случайная величина, имеющая нулевое математическое ожидание и дисперсию σ_ε^2, не зависящую от значений факторов $x_1,\dots,x_m$.

Характеристика $\Re^2 \equiv \Re^2(y; x_1,\dots,x_m)$, называемая *коэффициентом детерминации*, определяется формулой

$$\Re^2 \equiv 1 - \sigma_\varepsilon^2 / \sigma^2(y),$$

где $\sigma^2(y)$ — дисперсия случайной величины y. Показатель $\Re^2$ зависит от факторов $x_1,\dots,x_m$, но не от их значений, т.е. характеризует связь между y и факторами.

По данным выборки, состоящей из наблюдений $(\tilde{y}_k; \tilde{x}_{k1},\dots,\tilde{x}_{km}), k = 1,\dots,n$, с помощью МНК находятся оценки $\hat{a}_j$, $j = 0, 1,\dots,m$, коэффициентов в (1), МНК-остатки $\hat{\varepsilon}_k$, оценки $\sigma^2(\tilde{y}) = \frac{1}{n}\sum_k (\tilde{y}_k - \overline{\tilde{y}})^2$ и $\sigma^2(\tilde{y}; \tilde{x}_1,\dots,\tilde{x}_m) = \frac{1}{n}\sum_k \hat{\varepsilon}_k^2$ дисперсий $\sigma^2(y)$ и σ_ε^2, а затем выборочное значение $R^2(\tilde{y}; \tilde{x}_1,\dots,\tilde{x}_m) = 1 - \sigma^2(\tilde{y}; \tilde{x}_1,\dots,\tilde{x}_m) / \sigma^2(\tilde{y})$ коэффициента детерминации. При заданном наборе $\{x_1,\dots,x_M\}$ потенциальных факторов выбор набора $\{x_{i(1)},\dots,x_{i(m)}\}$ аргументов обычно сводят к минимизации несмещенной оценки $\frac{n}{n-p}\sigma^2(\tilde{y}; \tilde{x}_1,\dots,\tilde{x}_m)$ дисперсии σ_ε^2 или к максимизации статистики

$$\overline{R}^2(\tilde{y};\tilde{x}_1,...,\tilde{x}_m)\equiv\overline{R}^2\equiv R^2_{adj}\equiv 1-[1-R^2(\tilde{y};\tilde{x}_1,...,\tilde{x}_m)]\frac{n-1}{n-p}, \quad (2)$$

где $p=(m+1)$ — число оцениваемых коэффициентов в (1). **Статистику $\overline{R}^2$** называют *выборочным коэффициентом детерминации, скорректированным на число степеней свободы*.

Приведенные определения включаются в учебники по многомерному статистическому анализу и эконометрике. Менее известно, что математические ожидания ER^2 и $E\overline{R}^2$ статистик R^2 и $\overline{R}^2$ не равны $\Re^2$. Это важно, поскольку выбор факторов не должен быть ориентирован только на обеспечение наибольшей близости выравненных значений $\hat{y}_k=\hat{a}_0+\sum_{j=1}^{m}x_{kj}\hat{a}_j$, $k=1,\ldots,n$, переменной y к выборочным значениям y_k, так как уравнение $\hat{y}=\hat{a}_0+\sum_{j=1}^{m}x_j\hat{a}_j$ используется и при других значениях факторов. В связи с этим вводятся различные критерии выбора факторов, использующие предположения о генеральной совокупности переменных $y, x_1, \ldots, x_M$, например о нормальности соответствующего распределения. При таком предположении показатель $\Re^2$ естественно рассматривать как характеристику оцениваемой регрессии и, выбирая факторы, максимизировать ее несмещенную оценку.

2. Несмещенная оценка коэффициента детерминации $\Re^2$, ее аппроксимации и заменители

Уишарт [Wishart (1931)] показал, что $E\overline{R}^2$ и $\Re^2$ связаны соотношением [Кендалл, Стьюарт (1973), с. 454]

$$E\overline{R}^2=1-\frac{n-p}{n-1}(1-\Re^2)F(1;1;0{,}5(n+1);\Re^2). \quad (3)$$

Здесь $F(\alpha;\beta;\gamma;z)$ — специальная гипергеометрическая функция, определяемая в виде ряда [Градштейн, Рыжик (1962), с. 1053]

$$F(\alpha;\beta;\gamma;z)=1+\frac{\alpha\beta z}{1\cdot\gamma}+\frac{\alpha(\alpha+1)\beta(\beta+1)z^2}{1\cdot2\cdot\gamma(\gamma+1)}+\frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)z^3}{1\cdot2\cdot3\cdot\gamma(\gamma+1)(\gamma+2)}+\ldots, \quad (4)$$

сходящегося абсолютно и равномерно внутри единичного круга для комплексной переменной z, если $\gamma\neq0,-1,-2,\ldots$. Для дальнейшего важно, что функция $F(\alpha;\beta;\gamma;z)$ действительной переменной z при $z\geq0$ и положительных α,β,γ является возрастающей, а также то, что формула (3) не позволяет находить $E\overline{R}^2$ по данным выборки, так как связывает неизвестные детерминированные величины $\Re^2$ и $E\overline{R}^2$.

Важнейший результат был получен Олкиным и Прэттом [Olkin, Pratt (1958)], нашедшими определенную при $n>p\geq3$ **статистику $\widetilde{R}^2(\tilde{y};\tilde{x}_1,\ldots,\tilde{x}_m)\equiv\widetilde{R}^2$** [Кендалл, Стьюарт (1973), с. 456]:

$$\widetilde{R}^2=1-\frac{n-3}{n-p}(1-R^2)F(1;1;0{,}5(n-p)+1;1-R^2), \quad (5)$$

представляющую собой несмещенную оценку для $\Re^2(y;x_1,\ldots,x_m)$.

Свойства функции $F(1;1;\gamma;z)$ переменной z при $0 \le z \le 1$, $y = 0{,}5(n-p)+1$ известны: $F(1;1;\gamma;0)=1$; при $0 \le z < 1$ ряд (4) сходится, а при $z=1$ расходится, если $n-p=1$ или 2, и сходится, если $n-p \ge 3$ [Градштейн, Рыжик (1962), с. 1054].

Статистика $\widetilde{R}^2$ до настоящего времени, насколько нам известно, не использовалась, по-видимому, из-за признания практически невозможным или нецелесообразным вычислять значения $F(1;1;0{,}5q;z)$ при целых q и $0 < z < 1$.

В этих условиях можно воспользоваться аппроксимацией для $\widetilde{R}^2$, получаемой из (5) при большом числе наблюдений. В [Кендалл, Стьюарт (1973), с. 456] предлагается использовать первые члены разложения $\widetilde{R}^2$ в ряд

$$\widetilde{R}^2 = R^2 - \frac{p-3}{n-p}(1-R^2) - \frac{2(n-3)}{(n-p)(n-p+2)}(1-R^2)^2 - O(n^{-2}).$$

Таким образом, в рассмотрение вводится **статистика $\widetilde{\widetilde{R}}^2$**:

$$\widetilde{\widetilde{R}}^2 = R^2 - \frac{p-3}{n-p}(1-R^2) - \frac{2(n-3)}{(n-p)(n-p+2)}(1-R^2)^2, \qquad (6)$$

которую в [Айвазян и др. (1985), с. 284] предлагается применять как критерий качества регрессии. Из определений статистик $\overline{R}^2$ и $\widetilde{\widetilde{R}}^2$ следует, что при близких к нулю значениях R^2 они принимают отрицательные значения. Это же свойство отмечается в [Кендалл, Стьюарт (1973), с. 456–457] и для $\widetilde{R}^2$.

Сравним значения рассматриваемых статистик R^2, $\overline{R}^2$, $\widetilde{R}^2$ и $\widetilde{\widetilde{R}}^2$, не вычисляя их, но учитывая, что $0 < R^2 < 1$, $n > p \ge 3$ и $F \equiv F(1;1;0{,}5(n-p)+1;1-R^2) > 1$. Из (2), (5) и (6) получаем $R^2 - \overline{R}^2 = \dfrac{p-1}{n-p}(1-R^2) > 0$, $R^2 - \widetilde{\widetilde{R}}^2 > 0$, $\widetilde{\widetilde{R}}^2 - \widetilde{R}^2 > 0$, т. е.

$$\widetilde{R}^2 \le \max(\overline{R}^2; \widetilde{\widetilde{R}}^2) < R^2.$$

Покажем, что для статистик $\widetilde{\widetilde{R}}^2$ и $\overline{R}^2$ возможны случаи $\widetilde{\widetilde{R}}^2 < \overline{R}^2$, $\widetilde{\widetilde{R}}^2 > \overline{R}^2$ и $\widetilde{\widetilde{R}}^2 = \overline{R}^2$, и найдем множества значений величин n, p и R^2, при которых эти случаи имеют место.

Используя определения, представим разность этих статистик в виде

$$R^2 - \overline{R}^2 = 2(1-R^2)\frac{(n-p+2)(n-p+4) - (n-3)(1-R^2)[(n-p+4)+4(1-R^2)]}{(n-p)(n-3)(n-p+2)(n-p+4)}.$$

При фиксированных значениях n и p исследуем неопределенное неравенство

$$f(y) \equiv 4(n-3)y^2 + (n-3)(n-p+4)y - (n-p+2)(n-p+4) \vee 0,$$

в котором переменная $y = (1-R^2)$ удовлетворяет неравенству $0 \le y \le 1$. Очевидно, что уравнение $f(y)=0$ имеет корни y_-, y_+ разных знаков, неравенство $f(y) \le 0$ выполняется при $0 \le y \le \min(1; y)$. Имеем $\min(1; y_+) = y_+$, если $f(1) > 0$, но $\min(1; y_+) = 1$, если $f(1) \le 0$.

Таким образом, необходимо исследовать неравенство $f(1) = [4(n-3)+(n-3)(n-p+4) - -(n-p+2)(n-p+4)] \vee 0$, учитывая, что параметры n и p удовлетворяют условию $n > p \ge 3$. Введя неотрицательную переменную $x = (n-p-1) \ge 0$, представим неравенство $f(1) \vee 0$ в виде $(x+p-2)(x+9) - (x+3)(x+5) \equiv (p+1)(n-p+1) - 33 \vee 0$ или $n \vee [(p+1) + 33/(p+1)] \equiv \equiv h(p)$.

Рассмотрим три случая для пар (p; n) параметров, характеризующих регрессию, — числа наблюдений в выборке (n) и числа оцениваемых коэффициентов ($p = m + 1$).

1. Если $\boldsymbol{n \;\; h(p)}$, то $f(1)$ 0. Следовательно, $f(y) \equiv (\tilde{\bar{R}}^2 - \bar{R}^2) < 0$ при $0 \le (1 - R^2) < y_+$ и $\tilde{\bar{R}}^2 > \bar{R}^2$ при $y_+ < (1 - R^2) \le 1$, где y_+ — положительный корень уравнения $f(y) = 0$. Такие пары (p; n) будем называть **парами типа А.** Для них, т. е. для достаточно большого числа наблюдений, при больших значениях коэффициента детерминации R^2 скорректированный на число степеней свободы критерий $\bar{R}^2$ завышает оценку качества регрессии по сравнению с аппроксимирующим статистику $\Re^2$ критерием $\tilde{\bar{R}}^2$. Однако при малых R^2 такая оценка качеств а занижается.

2. При небольшом числе наблюдений n, удовлетворяющем неравенству $\boldsymbol{(p+1) \le n < h(p)}$, т. е. для **пар типа В**, имеем $f(1) < 0$ и $f(y) \equiv (\tilde{\bar{R}}^2 - \bar{R}^2) < 0$ для всех возможных значений R^2, т. е. при $0 < R^2 < 1$, и критерий $\bar{R}^2$ характеризует регрессию, преувеличивая оценку ее качества.

3. В особом случае, когда $\boldsymbol{n = h(p)}$, с учетом ограничения $n > p \ge 3$ существуют всего два значения $p = 10$ и $p = 32$, при которых $33/(p+1)$ и $h(p)$ — целые числа. Таким образом, $\tilde{\bar{R}}^2 = \bar{R}^2$ только при $p = 10$, $m = 9$, $n = 14$ или при $p = 32$, $m = 31$, $n = 34$, т. е. в двух исключительных и неинтересных для приложений случаях.

Для любого p значения n, образующие пары (p; n) этих типов, легко находятся. Так, при $p = 3$ А-множество пар (p; n) = (3; n) задается неравенством $12 \le n$, а В-множество представляется в виде (3; n), $n \in \{4; 5; 6; 7; 8; 9; 10; 11\}$. Например, при $p = 7$ такими множествами значений для n соответственно будут $12 \le n$ и $n \in \{8; 9; 10; 11\}$. Заметим, что при $p \ge 33$ А-множества задаются неравенствам и $(p + 2) \le n$, а В-множества «вырождаются» в $(p; n) \equiv (p; p + 1)$.

Приведенный анализ неравенства $\tilde{\bar{R}}^2 \le \max(\bar{R}^2; \tilde{\bar{R}}^2) < R^2$ показывает, что при $R^2 < 1$ статистики $\bar{R}^2$, $\tilde{\bar{R}}^2$, R^2 смещены относительно $\Re^2$ заведомо положительно, а для критериев $\bar{R}^2$ и $\tilde{\bar{R}}^2$ характер такого смещения зависит от параметров p, n и статистики R^2. Поэтому целесообразно продолжить поиск других подходов к конструированию на основе статистики R^2 критериев качества регрессий.

В [Айвазян и др. (1985), с. 190–192] было предложено при выборе регрессоров максимизировать не $\bar{R}^2$, а так называемую *нижнюю границу* $R^2_{\min,P}$ для $\Re^2$ *при задаваемой доверительной вероятности P*. Статистика $R^2_{\min,P}$ определялась при упрощающем предположении о пропорциональности разности $(R^2 - R^2_{\min,P})$ асимптотической (при больших n) оценке среднеквадратической ошибки случайной величины R^2. **Критерий $\boldsymbol{R^2_{\min,P}}$** задавался формулой

$$R^2_{\min,P} = R^2 - \lambda(P)(1 - R^2)\frac{2(p-1)(n-p)}{(n-1)^2(n+1)}. \qquad (7)$$

Значение множителя $\lambda(P)$ предлагалось выбирать в зависимости от P. Однако функция $\lambda(P)$ не поддается идентификации, и воспользоваться формулой (7) при ограниченном, а тем более при малом числе наблюдений невозможно. Значение коэффициента $\lambda(P)$ приходится задавать, исходя из прагматических соображений.

В развитие идеи, на которой основывалось введение статистики $R^2_{\min,\,p}$, в [Айвазян, Мхитарян (1998), с. 420, 663, 664] введен заменяющий статистику $\widetilde{R}^{\,2}$, просто вычисляемый, максимизируемый **показатель качества регрессии $R^2_{\min}$**:

$$R^2_{\min} = \overline{R}^{\,2} - 2(1-R^2)\left[\frac{2(p-1)(n-p)}{(n-1)^2(n+1)}\right]^{0,5}. \qquad (8)$$

Эта статистика также называется *нижней доверительной границей* (точнее, ее оценкой) для $\Re^2$, но без упоминания задаваемой доверительной вероятности.

Сравним значения статистик $\widetilde{R}^{\,2}$ и $R^2_{\min}$. Для разности $(\widetilde{R}^{\,2} - R^2_{\min})$, используя (5), (8) и неравенство $n > p \geq 3$, получаем

$$\widetilde{R}^{\,2} - R^2_{\min} = (1-R^2)\left\{\frac{(n-1)-(n-3)F}{n-p} + 2\left[\frac{2(p-1)(n-p)}{(n-1)^2(n+1)}\right]^{0,5}\right\},$$

где, как и прежде, $F \equiv F(1;1;0{,}5(n-p)+1;1-R^2)$. Из (5) находится следующая формула для $(n-3)F/(n-p)$: $(n-3)F/(n-p) = (1-\widetilde{R}^{\,2})/(1-R^2) \geq 1$. Тогда при $R^2 < 1$ имеем

$$\widetilde{R}^{\,2} - R^2_{\min} = (1-R^2)\left\{\frac{n-1}{n-p} - \frac{1-\widetilde{R}^{\,2}}{1-R^2} + 2\left[\frac{2(p-1)(n-p)}{(n-1)^2(n+1)}\right]^{0,5}\right\} > (1-R^2)\left\{\frac{n-1}{n-p} - 1 + 2\left[\frac{2(p-1)(n-p)}{(n-1)^2(n+1)}\right]^{0,5}\right\} =$$

$$= (1-R^2)\left\{\frac{p-1}{n-p} + 2\left[\frac{2(p-1)(n-p)}{(n-1)^2(n+1)}\right]^{0,5}\right\} > 0.$$

Следовательно, для математических ожиданий этих статистик имеем $E\widetilde{R}^{\,2} > ER^2_{\min}$ и статистика $R^2_{\min}$ смещена относительно $\Re^2$, что и следовало ожидать, учитывая их определения. В то же время из определений (5) и (8) для $\widetilde{R}^{\,2}$ и $R^2_{\min}$ следует, что с ростом n их значения сближаются, стремясь к R^2. Однако при ограниченном числе наблюдений эквивалентность применения критериев $R^2_{\min}$ и $\widetilde{R}^{\,2}$ в задаче выбора регрессий по меньшей мере не очевидна. Поэтому проанализируем возможность эффективного вычисления несмещенной оценки $\widetilde{R}^{\,2}$ для $\Re^2$.

3. Эффективно вычисляемая форма представления статистики $\widetilde{R}^2$

Чтобы оценка $\widetilde{R}^{\,2}$ для $\Re^2$ могла применяться в качестве критерия выбора множества регрессоров, достаточно иметь возможность вычислять значения функции $F(1;1;\gamma;z)$ при $\gamma = 0{,}5(n-p)+1$ и $0 \leq z = (1-R^2) \leq 1$. Это можно сделать следующими способами.

Во-первых, это значение можно рассчитывать, *используя определение* (4) для функции F. Тогда

$$\widetilde{R}^{\,2} = 1 - \frac{n-3}{n-p}(1-R^2)\left\{1 + (1-R^2)\sum_{k=0}^{\infty}\frac{k!(1-R^2)^k}{\gamma(\gamma+1)\cdots(\gamma+k)}\right\}. \qquad (5')$$

Однако такой способ может быть сложен для реализации из-за необходимости вычислять значения коэффициентов при $z^k = (1-R^2)^k$.

Во-вторых, можно воспользоваться представлением функции $F(1;1;\gamma;z)$ в виде *определенного интеграла* [Градштейн, Рыжик (1962), формула (9.111)]:

$$F(1;1;\gamma;z)=\frac{1}{\mathrm{B}(1;\gamma-1)}\int_0^1\frac{(1-u)^{\gamma-2}}{1-uz}\mathrm{d}u\equiv\frac{g(\gamma;z)}{\mathrm{B}(1;\gamma-1)}. \qquad (9)$$

Значение бета-функции $\mathrm{B}(1;\gamma-1)\equiv\mathrm{B}(1;0{,}5(n-p))$ легко вычисляется: $\mathrm{B}(1;\gamma-1)=\Gamma(1)\Gamma(\gamma-1)/\Gamma(\gamma)$, где $\Gamma(x+1)=\int_0^\infty e^{-t}t^x\mathrm{d}t$ — гамма-функция, $\Gamma(1)=1$, $\Gamma(x+1)=x\,\Gamma(x)$ и $\mathrm{B}(1;\gamma-1)=2/(n-p)$. Определенный интеграл $g(\gamma;z)$ может вычисляться методами численного интегрирования. Комбинируя формулы (5) и (9) и переходя к переменной $t=(1-u)$, получаем интегральное представление статистики $\widetilde{R}^2$:

$$\widetilde{R}^2=1-0{,}5(n-3)\int_0^1\frac{t^{0{,}5(n-p)-1}}{c+t}\mathrm{d}t, \qquad (5'')$$

где $c=R^2/(1-R^2)$ и $R^2\neq 1$.

Заметим, что с помощью (9) вычисляется значение $F(1;1;0{,}5(n-p)+1;1)$, получаемое при $R^2=0$, так как

$$g(\gamma;1)=\int_0^1(1-u)^{\gamma-3}\mathrm{d}u=\int_0^1 t^{\gamma-3}\mathrm{d}t \text{ и } \gamma-3=0{,}5(n-p-4).$$

Если $n=p+1$, то $\gamma-3=-1{,}5$, $\int t^{\gamma-3}\mathrm{d}t=-2t^{-0{,}5}$ и $g(\gamma;1)=+\infty$.

Если $n=p+2$, то $\gamma-3=-1$, $\int t^{\gamma-3}\mathrm{d}t=\ln t$ и $g(\gamma;1)=+\infty$.

Если $n\geq p+3$, то $\int_0^1 t^{\gamma-3}\mathrm{d}t=2/(n-p-2)$.

Таким образом, при минимальном значении $R^2=0$ статистики R^2 получаем

$$\widetilde{R}^2(0)\equiv 1-\frac{n-3}{n-p}F(1;1;0{,}5(n-p)+1;1)=\begin{cases}-\dfrac{p-1}{n-p-2} & \text{при } n\geq p+3;\\ -\infty & \text{при } n=p+1 \text{ или } p+2.\end{cases} \qquad (10)$$

В-третьих, функция $g(\gamma;z)$ при $\gamma=0{,}5(n-p)+1$ представима в виде *суммы конечного числа слагаемых, являющихся известными функциями аргументов* $(n-p)$ и $z=(1-R^2)$. Возможность получения такого представления до настоящего времени, по-видимому, не была замечена.

Для нахождения определенного интеграла в формуле (9) с параметром $(\gamma-2)=0{,}5(n-p)-1$, принимающим значения $\{-0{,}5;0;+0{,}5;1;\ldots\}$ при $n-p=1,2,\ldots$, введем переменную $z=(1-R^2)$. Предполагая, что $R^2<1$ и $c=(1-z)/z=R^2/(1-R^2)$, рассмотрим следующие случаи для $g(\gamma;z)$:

- При $n-p=2$ имеем

$$g(\gamma;z)=\frac{1}{z}\int_0^1\frac{1}{c+t}\mathrm{d}t=\frac{1}{z}\ln(1+c^{-1}).$$

• При нечетных значениях параметра $n-p=2s+1$, $s=1,2,\ldots$, применяя формулы (2.211) и (2.212) из [Градштейн, Рыжик (1962)], находим

$$g(\gamma; z)=\frac{1}{z}\int_0^1 \frac{t^{s-0,5}}{c+t}\mathrm{d}t=\frac{1}{z}\left[\sum_{k=0}^{s-1}\frac{c^k}{s-k-0,5}+(-1)^s\cdot 2c^{s-0,5}\mathrm{arctg}(c^{-0,5})\right].$$

• При четном $n-p=2(s+1)$, $s=1,2,\ldots$, используя формулу (2.153) из [Градштейн, Рыжик (1962)], получаем

$$g(\gamma; z)=\frac{1}{z}\int_0^1 \frac{t^s}{c+t}\mathrm{d}t=\frac{1}{z}\left[\sum_{k=0}^{s-1}\frac{c^k}{s-k}+(-1)^s c^s \ln(1+c^{-1})\right].$$

Приведенные формулы позволяют представить статистику $\widetilde{R}^2$ в виде

$$\widetilde{R}^2=1-0,5(n-3)G(n-p; c),$$

где функция $G(n-p; c)$ определена при $0<c=R^2/(1-R^2)$, $0<R^2<1$, $p=m+1$ следующим образом:

$$G(n-p; c)=2\sum_{k=0}^{s-1}(-1)^k\frac{c^k}{2(s-k)-1}+(-1)^s\cdot 2c^{s-0,5}\mathrm{arctg}(c^{-0,5})\equiv H_1(s; c)+H_2(s; c), \quad \text{если } n-p=2s+1;$$

$$G(n-p; c)=\sum_{k=0}^{s-1}(-1)^k\frac{c^k}{s-k}+(-1)^s c^s\ln(1+c^{-1})\equiv H_3(s; c)+H_4(s; c), \quad \text{если } n-p=2(s+1), \tag{11}$$

здесь $s=0, 1, 2,\ldots$.

Вычисление значений многочленов $H_1(s; c)$ и $H_3(s; c)$ от переменной c может приводить к потере точности при больших значениях c, т. е. при $R^2\approx 1$, поэтому для значений c и R^2 выделим следующие случаи:

1) $0\le R^2\le 0,5$, $c\le 1$, $c^{-1}\ge 1$;

2) $0,5\le R^2<1$, $c\ge 1$, $c^{-1}\le 1$.

При относительно небольших значениях R^2, т. е. при $R^2\le 0,5$, целесообразно использовать формулы (11) даже при больших значениях $(n-p)$, так как $c\le 1$. В случае $R^2\ge 0,5$ воспользуемся разложением функций arctg x и $\ln(1+x)$ в степенные ряды:

$$\mathrm{arctg}\,x=\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k+1}}{2k+1} \quad \text{при } x^2\le 1 \text{ и } x\equiv c^{-0,5}\le 1;$$

$$\ln(1+x)=\sum_{k=0}^{\infty}(-1)^k\frac{x^k}{k} \quad \text{при } -1<x\le 1 \text{ и } x\equiv c^{-1}\le 1. \tag{12}$$

С помощью (12) показывается, что многочлены $H_1(s; c)$ и $H_2(s; c)$ равны суммам слагаемых в функциях $H_3(s; c)$ и $H_4(s; c)$ соответственно, содержащих неотрицательные степени перемен-

ной c. Таким образом, для $G(n-p;\,c)$ при $R^2 \geq 0{,}5$ получаем общее для четных и нечетных значений $(n-p)$ представление в виде степенного ряда

$$G(n-p;\,c)=2c^{-1}\sum_{k=0}^{\infty}(-1)^k(2k+n-p)^{-1}c^{-k},$$

в котором $0 \leq c^{-1}=(1/R^2)-1 \leq 1$.

В итоге для статистики $\widetilde{R}^2$ получаем исчерпывающее все возможные случаи представление в виде функции от R^2, числа наблюдений n и числа коэффициентов p в уравнении регрессии (предполагается, что $n>p\geq 3$, $c=R^2/(1-R^2)$ и q — целая часть числа $0{,}5(n-p-1)$):

$$\widetilde{R}^2=\begin{cases}-\infty, & \text{если } R^2=0 \text{ и } n=p+1,\,p+2;\\ -\dfrac{p-1}{n-p-2}, & \text{если } R^2=0 \text{ и } n\geq p+3;\\ 1-(n-3)\left[\displaystyle\sum_{k=0}^{q-1}(-1)^k\frac{c^k}{n-p-2k-2}+(-1)^q c^{q-0,5}\operatorname{arctg}(1/R^2-1)^{0,5}\right], & \text{если } 0<R^2<1 \text{ и } n-p=2q+1;\\ 1-(n-3)\left[\displaystyle\sum_{k=0}^{q-1}(-1)^k\frac{c^k}{n-p-2k-2}+(-1)^q c^q \ln(1/R^2)^{0,5}\right], & \text{если } 0<R^2<1 \text{ и } n-p=2(q+1);\\ 1, & \text{если } R^2=1.\end{cases} \quad (5''')$$

В случае если $0{,}5 \leq R^2<1$ и $c\geq 1$, можно также воспользоваться формулой

$$\widetilde{R}^2=1-(n-3)c^{-1}\sum_{k=0}^{\infty}(-1)^k(2k+n-p)^{-1}c^{-k}. \quad (5'''')$$

Представление (5'''), (5'''') статистики $\widetilde{R}^2$ по сравнению с (5'), использующим коэффициенты $k!/[\gamma(\gamma+1)\cdots(\gamma+k)]$ при $(1-R^2)^k$, отличается простотой формул для коэффициентов при степенях переменных c и c^{-1}. Метод вычисления статистики $\widetilde{R}^2$ по формулам (5''') реализован совместно с канд. экон. наук Н. А. Толмачевой.

4. Примеры применения статистики $\widetilde{R}^2$ при выборе наилучшей регрессии

Подходы к выбору наилучшей регрессии в задаче с небольшим числом потенциальных факторов в научных монографиях и учебниках иллюстрируются на нескольких повторяемых примерах. Это позволяет сравнивать результаты, получаемые с использованием постоянно обновляемых идей и общих исходных данных. На двух таких примерах продемонстрируем возможность применения статистики $\widetilde{R}^2$.

4.1. Пример Хальда

В [Дрейпер, Смит (1987)] и [Себер (1980)] детально анализируются все варианты регрессий, базирующихся на данных из [Woods et al. (1932)] и [Хальд (1956)]. Объясняемая переменная $y=(y_k)$ в этом примере — тепло, выделяющееся при производстве цемента (калория/грамм), а $x_j=(x_{kj})$, $j=1,\ldots,4$ $(m=4)$, — переменные, характеризующие содержание четырех веществ в клинкере (в %) в 13 наблюдениях ($k=1,\ldots,n; n=13$). Факторы x_j приближенно мультиколлинеарны, так как их суммы в каждом наблюдении близки к 100. Вы-

борочные значения коэффициентов корреляции для пар факторов подтверждают предположение о мультиколлинеарности ($r(x_1; x_3) \cong -0{,}8241$, $r(x_2; x_4) \cong -0{,}9730$), так же как и значение $\det(\mathbf{X'X}) \cong 0{,}0010677$ детерминанта матрицы $\mathbf{X'X}$, где $\mathbf{X}$ — матрица размером 13×5, содержащая значения аргументов в регрессии $y = a_0 + a_1 x_1 + \ldots + a_4 x_4$, и собственные значения корреляционной матрицы $\mathbf{C} \equiv \mathrm{cor}(x_1, \ldots, x_4)$ для факторов: $\lambda_1(\mathbf{C}) \cong 2{,}23569$, $\lambda_2(\mathbf{C}) \cong 1{,}57606$, $\lambda_3(\mathbf{C}) \cong 0{,}18661$ и $\lambda_4(\mathbf{C}) \cong 0{,}00162$.

С использованием различных подходов в [Дрейпер, Смит (1987)] и [Себер (1980)] были выделены следующие претенденты на роль набора факторов для наилучшей регрессии:

$$(x_1; x_2),\ (x_1; x_4),\ (x_1; x_2; x_3),\ (x_1; x_2; x_4),\ (x_1; x_3; x_4),\ (x_2; x_3; x_4),\ (x_1; x_2; x_3; x_4).$$

Таблица 1 содержит значения статистик R^2, $\overline{R}^2$, $R^2_{\min}$, $\widetilde{\widetilde{R}}^2$ и $\widetilde{R}^2$ для всех 15 вариантов набора факторов $x_1, \ldots, x_4$. В этом примере значения статистик $\widetilde{\widetilde{R}}^2$ и $\widetilde{R}^2$ приводятся с бóльшим числом знаков для того, чтобы сделать явным выполнение неравенства $\widetilde{\widetilde{R}}^2 > \widetilde{R}^2$. Отобранные варианты четко выделяются среди регрессий с фиксированным числом факторов. При этом регрессии с одним фактором ($m = 1$, $p = 2$) уступают по критериям $\overline{R}^2$ и $R^2_{\min}$ регрессиям-претендентам.

Таблица 1

Значения критериев выбора регрессии, основанных на функциях от статистики R^2, для примера Хальда

Набор факторов	Статистика R^2	Максимизируемые критерии				Ранг набора факторов*
		$\overline{R}^2$	$R^2_{\min}$	$\widetilde{\widetilde{R}}^2$	$\widetilde{R}^2$	
(x_1)	0,53395	0,49158	0,39421	—	—	—
(x_2)	0,66627	0,63593	0,56620	—	—	—
(x_3)	0,28587	0,22095	0,07175	—	—	—
(x_4)	0,67454	0,64495	0,57696	—	—	—
$(x_1; x_2)$	0,97868	0,97441	0,96841	0,9786026	0,9786021	4
$(x_1; x_3)$	0,54817	0,45780	0,33051	0,5141412	0,5088098	11
$(x_1; x_4)$	0,97247	0,96697	0,95921	0,9723448	0,9723437	6
$(x_2; x_3)$	0,84703	0,81643	0,77333	0,8431252	0,8429443	9
$(x_2; x_4)$	0,68006	0,61607	0,52594	0,6630002	0,6612219	10
$(x_3; x_4)$	0,93529	0,92235	0,90412	0,9345918	0,9345785	8
$(x_1; x_2; x_3)$	0,98228	0,97638	0,97058	0,9802529	0,9802526	2
$(x_1; x_2; x_4)$	0,98234	0,97645	0,97067	0,9803097	0,9803094	1
$(x_1; x_3; x_4)$	0,98128	0,97504	0,96891	0,9791304	0,9791300	3
$(x_2; x_3; x_4)$	0,97282	0,96376	0,95486	0,9696507	0,9696495	7
$(x_1; x_2; x_3; x_4)$	0,98238	0,97356	0,96728	0,9778919	0,9778914	5

* Приведены ранги регрессий, для которых определены статистики $\overline{R}^2$, $R^2_{\min}$, $\widetilde{\widetilde{R}}^2$ и $\widetilde{R}^2$. Ранги присваиваются в соответствии с убыванием значений любого из критериев.

Для регрессии с факторами $(x_1; x_2)$ значения статистик $\overline{R}^2$, $R^2_{\min}$, $\widetilde{\overline{R}}^2$ и $\widetilde{R}^2$ больше, чем для регрессии с факторами $(x_1; x_4)$. Аналогичным образом регрессия с факторами $(x_1; x_2; x_4)$ оказывается предпочтительнее других регрессий с тремя и двумя факторами. Дрейпер и Смит, используя *метод исключения факторов* и «*шаговый метод*» (метод пополнения множества факторов), принимая без тестирования гипотезу нормальности ошибок и задавая без обоснования уровень значимости для *F*-критериев, отдали предпочтение регрессии с факторами $(x_1; x_2)$. В качестве критерия выбора факторов ими использовалась и предложенная Мэллоузом C_p-*статистика*, что также привело к выбору регрессии с факторами $(x_1; x_2)$. Однако при этом не было обращено внимание на то, что в этом критерии в качестве надежной, по предположению несмещенной оценки дисперсии случайных ошибок используется такая величина, как «s^2 — остаточный средний квадрат МНК-отклонения для уравнения, содержащего все переменные» [Дрейпер, Смит (1987), с. 14, 15]. Для примера Хальда с явно мультиколлинеарными данными указанное допущение вряд ли может быть оправдано. Такой оценкой было бы естественнее считать статистику s^2 для искомой «наилучшей регрессии», но это разрушало бы конструкцию метода, использующего статистику C_p.

Полезно иметь в виду, что так называемая *ПРЕСС-процедура* [Дрейпер, Смит (1987), с. 40–42] тоже позволила выделить варианты регрессий, для которых критерий «предсказанная сумма квадратов» (Prediction sum square) $\mathrm{PSS}(x_{j(1)}, \ldots, x_{j(m)})$ принимал наименьшие, но относительно мало различающиеся значения: $\mathrm{PSS}(x_1; x_2) \cong 95$, $\mathrm{PSS}(x_1; x_4) \cong 121$, $\mathrm{PSS}(x_1; x_2; x_3) \cong 91$, $\mathrm{PSS}(x_1; x_2; x_4) \cong 85$, $\mathrm{PSS}(x_1; x_3; x_4) \cong 87$, $\mathrm{PSS}(x_1; x_2; x_3; x_4) \cong 110$. Для остальных регрессий значения критерия PSS оказались в пределах от $\mathrm{PSS}(x_3; x_4) \cong 264$ до $\mathrm{PSS}(x_3) \cong 2616$. По-видимому, стремление выбирать уравнение как можно с меньшим числом аргументов хотя бы частично объясняется преувеличением трудностей реализации МНК, возникающих с ростом числа факторов. Однако для регрессий с двумя и тремя факторами эта позиция авторов не может объясняться возрастающей «сложностью» расчетов. Скорее следовало бы говорить об угрозе возникновения мультиколлинеарности факторов с увеличением их числа и о необходимости прогнозировать большее число факторов.

Можно считать, что в данном примере ПРЕСС-процедура в качестве конкурирующих регрессий определяет уравнения с наборами факторов $(x_1; x_2; x_4)$, $(x_1; x_3; x_4)$, для которых значения критерия PSS минимальны. При этом в число конкурирующих претендентов включена регрессия $(x_1; x_2; x_4)$ с наибольшими значениями статистик $\overline{R}^2$, $R^2_{\min}$, $\widetilde{\overline{R}}^2$ и $\widetilde{R}^2$.

Этот же набор факторов $(x_1; x_2; x_4)$ определяется в качестве наилучшего и при применении предложенного в [Webster et al. (1974)] *модифицированного МНК*, или *метода* «*регрессии на главные компоненты*». Этот метод использует собственные векторы корреляционной матрицы для объясняемой переменной и всех рассматриваемых факторов. Формальное изложение метода и его применение к данным примера Хальда имеются в [Дрейпер, Смит (1987), с. 48–52].

Себер, используя понятие R^2-адекватного (α)-набора регрессоров, предложенное в [Aitkin (1974)], приводит все такие наборы для примера Хальда, соответствующие доверительной вероятности $\alpha = 0{,}05$. Ими оказались $(x_1; x_2)$, $(x_1; x_4)$ и все четыре набора, содержащие три фактора [Себер (1980), с. 351, 352]. Однако этот подход не позволил в этом примере сузить множество регрессий-конкурентов.

Несовпадение результатов выбора наилучшей регрессии разными методами или фактическая неединственность результатов такого выбора отмечается почти всеми исследовате-

лями. Так, в [Себер (1980), с. 372] замечено, что метод последовательного включения факторов выделяет набор $(x_1; x_2; x_4)$, в то время как метод последовательного их исключения — набор $(x_1; x_2)$. Заметим, что в этих методах доверительные вероятности задаются экзогенно, без учета того, насколько различаются значения возможных критериев качества регрессий по наборам факторов, и без тестирования нормальности.

Таким образом, рассматриваемые Дрейпером, Смитом и Себером методы определения наилучшей регрессии в примере Хальда фактически позволили выделить множество регрессий-конкурентов, а не одну, действительно лучшую, регрессию.

В то же время на примере Хальда видно, что для вариантов регрессий со значениями R^2, близкими к 1, статистики $\widetilde{R}^2$ и $\widetilde{\widetilde{R}}^2$ становятся, как отмечалось, почти равными. В этом примере ранги, присвоенные регрессиям по убыванию значений критериев $\overline{R}^2$, $R^2_{\min}$, $\widetilde{\widetilde{R}}^2$ и $\widetilde{R}^2$, не являющихся неубывающими при добавлении факторов, совпадают. Следовательно, применение несмещенной оценки $\widetilde{R}^2$ для коэффициента детерминации $\Re^2$ как критерия качества регрессий в этом случае не противоречит рекомендациям применять другие рассматриваемые критерии.

4.2. Анализ урожайности зерновых культур

По данным 20 сельскохозяйственных районов некоторой области в примере 15.1 из [Айвазян, Мхитарян (1998), с. 631, 632, 636, 644–646, 652, 654, 664–668] исследуется зависимость урожайности зерновых культур y (ц/га) от пяти факторов: x_1 — число тракторов на 100 га; x_2 — число зерноуборочных комбайнов на 100 га; x_3 — число орудий поверхностной обработки почвы на 100 га; x_4 — количество удобрений, расходуемых на гектар (ц/га); x_5 — количество расходуемых химических средств защиты растений (ц/га). Отмечается высокая мультиколлинеарность факторов, причем коррелированность факторов x_1 и x_3 следует из того, что «орудия поверхностной обработки почвы реализуются в подавляющем большинстве с помощью тракторов» [Айвазян, Мхитарян (1998); с. 652, 654]. Поэтому из дальнейшего анализа исключим фактор x_1.

В табл. 2 приведены значения статистик R^2, $\overline{R}^2$, $R^2_{\min}$, $\widetilde{\widetilde{R}}^2$ и $\widetilde{R}^2$ для всех вариантов регрессий. Среди уравнений с одним фактором ($m=1$, $p=2$) явно выделяется регрессия с фактором x_4, для которой значения всех рассчитанных критериев существенно превосходят их значения для других однофакторных уравнений. Из множества уравнений с двумя факторами ($m=2$) по значениям всех пяти статистик выделяются регрессии с факторами $(x_2; x_4)$ и $(x_3; x_4)$. Для уравнения с факторами $(x_3; x_4)$ значения всех максимизируемых статистик больше, чем для регрессии с факторами $(x_2; x_4)$. Среди трехфакторных регрессий по значениям всех статистик претендентами на роль наилучшей регрессии оказываются уравнения с наборами факторов $(x_2; x_4; x_5)$ и $(x_3; x_4; x_5)$. Однако для регрессии с факторами $(x_2; x_4; x_5)$ значения статистик больше, чем у конкурирующего уравнения. Таким образом, выбор наилучшей регрессии сводится к выбору между уравнениями с факторами $(x_3; x_4)$ и $(x_2; x_4; x_5)$, поскольку для «лучшей» однофакторной регрессии значения статистик R^2, $\overline{R}^2$ и $R^2_{\min}$ существенно меньше, чем для этих претендентов. Напомним, что для регрессий с одним фактором не все рассматриваемые статистики определены. Регрессия с четырьмя факторами уступает отобранным двум конкурирующим уравнениям по всем критериям за исключением R^2, что естественно.

Таблица 2

Значения критериев выбора регрессии, основанных на функциях от статистики R^2, для примера анализа урожайности зерновых культур

Набор факторов	Статистика R^2	Максимизируемые критерии				Ранг набора факторов*
		$\overline{R}^2$	$R^2_{\min}$	$\widetilde{\widetilde{R}}^2$	$\widetilde{R}^2$	
(x_2)	0,13994	0,09215	–0,02638	—	—	—
(x_3)	0,16253	0,11601	0,00058	—	—	—
(x_4)	0,33329	0,29625	0,20436	—	—	—
(x_5)	0,11031	0,06089	–0,06173	—	—	—
(x_2; x_3)	0,16408	0,06573	–0,09261	0,09052	0,07524	11
(x_2; x_4)	0,46196	0,39866	0,29674	0,43148	0,42783	4
(x_2; x_5)	0,17248	0,07512	–0,08163	0,10039	0,08562	10
(x_3; x_4)	0,48237	0,42147	0,32342	0,45416	0,45093	2
(x_3; x_5)	0,21503	0,12268	–0,02601	0,15017	0,13776	8
(x_4; x_5)	0,33330	0,25486	0,12858	0,28651	0,27924	7
(x_2; x_3; x_4)	0,48386	0,38708	0,27092	0,42015	0,41635	6
(x_2; x_3; x_5)	0,22120	0,07518	–0,10010	0,10093	0,08651	9
(x_2; x_4; x_5)	0,51346	0,42223	0,31273	0,45510	0,45195	1
(x_3; x_4; x_5)	0,49823	0,40415	0,29122	0,43715	0,43367	3
(x_2; x_3; x_4; x_5)	0,51730	0,38858	0,26712	0,42188	0,41819	5

* Приведены ранги регрессий, для которых определены статистики $\overline{R}^2$, $R^2_{\min}$, $\widetilde{\widetilde{R}}^2$ и $\widetilde{R}^2$. Ранги присваиваются в соответствии с убыванием значений любого из критериев.

С. А. Айвазян и В. С. Мхитарян, рекомендующие статистику $R^2_{\min}$ как критерий качества регрессии, отдают предпочтение уравнению с факторами (x_3; x_4), так как $R^2_{\min}(\widetilde{y}\,;\,\widetilde{x}_3\,,\,\widetilde{x}_4) \cong 0{,}323 > 0{,}313 \cong$ $\cong R^2_{\min}(\widetilde{y}\,;\,\widetilde{x}_2\,,\,\widetilde{x}_4\,,\,\widetilde{x}_5)$. Однако по значениям статистик $\overline{R}^2$, $\widetilde{R}^2$ и $\widetilde{\widetilde{R}}^2$ регрессия с факторами (x_2; x_4; x_5) предпочтительнее, хотя разницы значений критериев для этих двух конкурирующих уравнений малы. Таким образом, на данном примере показано, что выбор регрессии по критериям $R^2_{\min}$ и $\widetilde{R}^2$ может приводить к разным результатам. Значения статистик $\widetilde{R}^2$ и $\widetilde{\widetilde{R}}^2$ могут для данного набора факторов существенно различаться, но при этом ранги регрессий, присваиваемые в соответствии с убыванием этих критериев, могут полностью или частично совпадать.

5. Заключение

Предложение использовать несмещенную оценку $\widetilde{R}^2$ коэффициента детерминации $\Re^2$ или ее аппроксимацию $\widetilde{\widetilde{R}}^2$ как критерий качества выбираемого набора регрессоров основывается на строго формулируемом предположении о нормальности распределения для совокупности переменных, порождающих используемые выборочные данные, и на теоретическом определении показателя качества зависимости одной из таких переменных от заданного набора других переменных-факторов. При применении статистики $\widetilde{R}^2$ не используется

предположение о большом числе наблюдений. В этом состоят преимущества предложенного подхода к определению конкурирующих регрессий по сравнению с эвристическими по своему характеру методами, использующими статистики $\overline{R}^2$ и $R^2_{\min}$. Реализованный метод расчета значений критерия-статистики $\widetilde{R}^2$ универсален и эффективен в широком диапазоне целочисленных характеристик уравнений регрессии — числа наблюдения и числа оцениваемых коэффициентов.

То, что в рассмотренных примерах применение статистики $\widetilde{R}^2$ приводит к выделению наборов регрессоров, полученных другими, более простыми в реализации методами, может рассматриваться как оправдание использования эвристических методов в конкретных случаях, но не означает эквивалентность таких методов в общем случае.

Поскольку статистика $\widetilde{R}^2$ и другие сравниваемые статистики представляют собой случайные величины, можно считать, что их применение как критериев качества наборов факторов в регрессии с общей выбранной объясняемой переменной позволяет всего лишь выделять конкурирующие варианты регрессий, для которых значения критериев близки. Выбор предпочтительных вариантов регрессий из множества конкурирующих, а в перспективе и конструирование с использованием отобранных регрессий уравнений, моделирующих объясняемую переменную, по-видимому, можно и целесообразно основывать на специально обсуждаемых качественных требованиях к ним. Обоснование таких конструктивно реализуемых требований — задача проводимых в настоящее время исследований.

Список литературы

Айвазян С. А., *Енюков И. С.*, *Мешалкин Д. Д.* Прикладная статистика. Исследование зависимостей: Справочное издание. М.: Финансы и статистика, 1985.

Айвазян С. А., *Мхитарян В. С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.

Градштейн И. С., *Рыжик И. М.* Таблицы интегралов, сумм, рядов и произведений. М.: Гос. изд. физ.-мат. литературы, 1962.

Дрейпер Н., *Смит Г.* Прикладной регрессионный анализ. Книга 2. М.: Финансы и статистика, 1987.

Кендалл М., *Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973.

Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980.

Хальд А. Математическая статистика с техническими приложениями. М.: ИЛ, 1956.

Aitkin M. A. Simultaneous inference and the choice of variable subsets // *Technometrics*. 1974. V. 16, P. 221–227.

Olkin I., *Pratt J. W.* Unbaised estimation of certain correlation coefficients // *Ann. Math. Statist*. 1958. V. 29.

Webster J. T., *Gunst R. F.*, *Mason R. L.* Latent root regression analysis // *Technometrics*. 1974. V. 16. P. 513–522.

Wishart J. The mean and second moment coefficient of the multiple correlation coefficient in sumples from a normal population // *Biometrica*. 1931. V. 22.

Woods H., *Steinour Y. H.*, *Starke H. R.* Effect of Composition of Portland on Heat Evolved during Hardening // *Industrial and Engineering Chemistre*. 1932. V. 24. P. 1207–1214.